LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-222460

Adapting Existing Spatial Data Sets to New Uses: An Example from Energy Modeling

G. Johanesson, J. S. Stewart, C. Barr, L. B. Sabeff, R. George, D. Heimiller, A. Milbrandt

June 27, 2006

Journal of Map & Geography Libraries

Updated October 14, 2003

Adapting Existing Spatial Data Sets to New Uses: An Example from Energy Modeling

Gardar Johannesson, Jeffrey Stewart, Christopher Barr

Lawrence Livermore National Laboratory

And

Liz Brady Sabeff, Ray George, Donna Heimiller, Anelia Milbrandt

National Renewable Energy Laboratory

Abstract.

Energy modeling and analysis often relies on data collected for other purposes such as census counts, atmospheric and air quality observations, and economic projections. These data are available at various spatial and temporal scales, which may be different from those needed by the energy modeling community. If the translation from the original format to the format required by the energy researcher is incorrect, then resulting models can produce misleading conclusions. This is of increasing importance, because of the fine resolution data required by models for new alternative energy sources such as wind and distributed generation. This paper addresses the matter by applying spatial statistical techniques which improve the usefulness of spatial data sets (maps) that do not initially meet the spatial and / or temporal requirements of energy models. In particular, we focus on (1) aggregation and disaggregation of spatial data, (2) imputing missing data and (3) merging spatial data sets.

1. Introduction.

Domestic coal, natural gas, and nuclear technology made up 88% of U.S. electricity production in 2003. These resources are generated by dispatchable plants, so estimating average power

production, capital requirements and operating expenses fits neatly within traditional cost and economic forecast models. Furthermore, these models require statistical techniques which are currently common within the energy modeling community.

However, new energy sources, such as wind turbines, solar photovoltaic, biomass, and electrical storage have much more complex spatial and temporal data requirements. The timing and magnitude of their power production do not follow the same patterns as dispatchable generation, and it is often more difficult to estimate their fuel supplies.

Of primary concern is the National Energy Modeling System (NEMS), which is maintained by the U.S. Department of Energy and the Energy Information Agency. NEMS is particularly important, because of its critical role in policy decisions. Developed in 1993, the 11 modules and 13 regions of NEMS reflect the spatial and temporal scales of the United States' twentieth century energy system. However, increasing public interest in alternative energy – for example, as shown by Renewable Portfolio Standards in several states, and increasing federal support for renewable energy research – has made intermittent technologies more important.

2. Future Energy Technologies Require Fine Resolution Data for Analysis.

Unlike dispatchable generators under the control of operators, intermittent technologies have capacity factor patterns which are irregular. In the reporting process these data are commonly aggregated into "typical" energy demand periods which can introduce errors into the final analysis if proper spatial statistical techniques are not included in the analysis.

Example One: Altamont Wind Production and Energy Demand

As an example, Lamont and Wu (2005) consider data for hourly energy demand and wind production from the Altamont wind site in northern California. The original data set is quite large, including nearly 9000 records, and for display purposes, a representative 10 day sample is used. The two images in Figure 1 illustrate the potential problems of aggregation. Figure 1(a) shows that actual wind generation data are highly irregular, but the normalized electricity demand data are rather stable. If wind generators were the only producers, the system shows alternating patterns of excessive and inadequate energy production. The results of aggregating data for this period are shown in Figure 1(b). The wind data have been represented by a continuous production pattern, indicating wind resources never reach zero or exceed demand. In this example, aggregation results in important information being lost and leads to the incorrect conclusion that wind resources will always be generating some power and will therefore always be used. End Example One.

Figure 1 Approximately Here

Example Two: Wind Power Density

A second example is a map of wind power density (provided to LLNL by TrueWind Solutions) covering approximately a 9000 square kilometer region near the San Francisco Bay Area. Wind power density at a height of 50m was originally produced at a 200m resolution shown in Figure 2 along with five LLNL produced aggregated scales. Figure 3 shows a histogram produced by LLNL of the wind-power density for all six resolutions. While the wind - scales 400-500 (class

4) and 500-600 (class 5) – are of significant importance to energy modelers, we see a critical amount of information is lost with increasing aggregation. The amount of class 5 wind is actually .21% at the original 200m resolution, but only .11% at the 800-m resolution level. An energy modeler using the coarser resolution would come up with a very different result than one using the finer resolution. The use us a histogram can inform the energy modeler which is the appropriate resolution to use. End Example Two.

Figure 2 Approximately Here

Figure 3 Approximately Here

Thus, the ability to aggregate data depends on the level of bias being introduced. That will depend on many factors in the analysis, such as whether time-of-day pricing or fixed pricing is used and whether the wind production is higher during the peak hours and lower during the off-peak hours. If all important factors are not understood, bias can be introduced, causing errors in the final energy analysis.

To determine capacity and cost factors, intermittent technologies often require information and data at a relatively fine spatial and temporal scale because of their (uncontrollable) natural variability. However, such information might not necessarily exist at the scales needed but may be available at courser scales gathered for an altogether different purpose.

Example Three: Coarse Resolution Biomass Data.

The United States Department of Agriculture (USDA) currently collects the desired biomass data, but at the county level. For energy modeling purposes, such as accurate cost analysis, it is important to model transportation distances and transmission line proximity to various sites of interest with high precision.

If county wide data are used, analysts are forced to simply put each county's measurement at a single point, or assume it is spread uniformly over the county. In Section 4 we compare these methods with statistical techniques which use secondary information, such as land use or population data, to produce finer resolution biomass maps.

3. Spatial Analysis and Statistical Modeling.

Spatial data are commonly recorded on points (called point-referenced data) or small regions (called areal data). Well established statistical techniques exist for analyzing both types. In this section we will briefly introduce a few of the techniques used in this paper. We refer to Cressie (1993), Banerjee, Carlin, and Gelfand (2004), Johannesson and Stewart et al. (2006) and Schabenberger and Gotway (2005) for further details.

Given a coarse-resolution spatial map, it is often of interest to interpolate the data on a finer-resolution map. Similarly, given point-reference data, it is often of interest to interpolate (predict) at unobserved location, or even interpolate on a fine-resolution map. The general method we employ in this paper is spatial regression. In this technique, data are thought to be realizations of a spatial process. Within this framework, each observed data point or areal-unit (pixel) is modeled as a combination of the large scale spatial trend, the small scale variation, and

the error in the data. Each of these three parts is modeled separately. The large scale trend can include information related to the variable of interest (for example, our large scale trend in Section 4 includes land use data). The small scale variation term introduces spatial correlation to the model, to leverage the fact that nearby data tend to be alike. And the data error accounts for possible uncertainty in the data, for example, measurement errors in point-referenced data. The general framework outlined here applies to both point-referenced and areal data, and can be used to solve problems were data are misaligned or need to be disaggregated.

When building these models it is important to have a familiarity with the raw data. One helpful tool is the semivariogram. The semivariogram explores similarity between points as a function of distance and is therefore a vital tool in spatial modeling. In addition, because the semivariogram gives us a sense of how smooth the data are, it can be helpful in determining how fine the resolution of the data must be.

4. Application: Biomass Data

Recall the crop residue biomass data introduced in Example Three in Section 2 Figure 4(a) shows crop residue biomass by county. Given this information, our goal is to conduct a transportation feasibility study. While basic techniques for handling the data exist (such as placing all data at the county's centroid or evenly distributing the data over the county) these are clearly not ideal. Our goal then is to employ spatio-statistical techniques using the correlated land use data (Figure 4(c)) to produce more accurate results.

Figure 4 Approximately Here

The first step in our analysis is determining the sufficient spatial scale for representing the underlying biodensity process. Figure 5 shows the empirical semivariogram of the biodensity using the distance between county centroids as a course proxy for "distance" between counties. As expected the semivariogram shows strong spatial correlation due to the smooth variation in the biodensity Figure 4(b). Given this consideration only, such smooth variation suggest that we can represent the biodensity at a relatively course scale. However, our aim is to answer questions about proximity, and as such, distances at the county-level are too large. We therefore aim to map the biomass data on a finer-resolution spatial grid for more accurate proximity analysis. With these considerations in mind, we have selected a 4km grid point spacing for this analysis, yielding 13,663 grid points that cover Minnesota.

Figure 5 Approximately Here

Our first spatial model does not take advantage of land use data and simply consists of a smooth large-scale spatial trend component and a small-scale spatial variation component . A spatial kernel smoother (locally weighted average) was applied to the county biodensity data (assigned to the centroid of each county) to extract the smooth spatial trend, which is shown in Figure 6(a). Once the trend had been extracted, the small-scale spatial variation component was fitted (using a spherical spatial correlation function with variation proportional to the extracted trend). Figure Figure 6(b) shows the resulting interpolated (predicted) residue biomass density map along with

a prediction uncertainty map Figure 6(c). On a final note, the spatial model was constrained to "honor" the data in the sense that when the pixel map of Figure 4(a) is aggregated up to the county-level, it yield the original county-level data.

Figure 6 Approximately Here

For our second model, we incorporated land-use data into the model as well. Land-use data were reported at approximately 1km resolution. We suspect crop related land use may yield crop residue, and within Minnesota, there are three major crop related land-use types (Cropland / Pasture, Cropland / Grassland, Cropland / Woodland). Our estimated biomass per pixel is determined by a combination of these three land-use categories and small scale variation. Figure Figure 7 shows the land-use trend, the interpolated (predicted) biomass density map, and a prediction uncertainty map.

Figure 7 Approximately Here

To compare the results of our models and existing techniques, three locations were selected – one surrounded by homogeneous, high-yield crop land; one near and urban area; and one in an area of variable residue density. A comparison of the amount of residue biomass as a function of

distance is presented in Figure 8 for the three locations. There is remarkably little difference between assuming that the biomass is evenly distributed versus the spatial models developed. However, putting the biomass at the centroid of each county shows deviation from the other three models. The results for the various methods are largely similar, because the typical size of a county is well within the range of the spatial correlation, as is evident in Figure 5. Less smooth data would have resulted in a larger difference between sophisticated and naïve methods.

Figure 8 Approximately Here

Discussion

Many spatial and spatiotemporal modeling techniques can be very useful in assisting energy modelers to extract resource information, demand data, and cost factors needed by merging information from various, disparate sources. In addition to those mentioned in this paper, Johannesson and Stewart et al. (2005) discuss other methods for working with point, areal and misaligned data, by modeling the unknown spatial process, which are also highly applicable to the data needs of energy modelers.

Additionally, spatial and spatiotemporal modeling continues to be an active field of research within statistics. An important, recent development is the use of hierarchical Bayesian methods, which exploit sampling-based techniques. These methods differ from more traditional methods

in that results are presented as an ensemble of possible outcomes (for example, possible maps given available data) instead of a static result.

Techniques developed by the spatial modeling community can be applied to the energy modeling area for use in analyzing resource information and new technologies that require more robust methods of estimation. Energy modelers need to have a sense of the importance of spatial and temporal scales required for accurate representation of the energy system under review. This information is important for the spatial modeling team's ability to determine which statistical procedures to apply. There are numerous ways of merging data sets that exist at different resolutions, and an equal number of tests to validate the results. The type and number of techniques to apply will depend on the level of precision required by the energy modelers to support their analysis. The spatial and energy modeling teams need to exchange information on (1) the questions energy modelers are trying to answer (2) the spatial and temporal scales desired (3) the data sets available and (4) the minimum acceptable accuracy. This exchange will allow the spatial modeling team to determine the best procedures and tests to validate results.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

References:

Banerjee, S., B.P. Carlin and A.E. Gelfand. 2004. *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman and Hall/CRC.

Cressie, N.A.C. 1993. *Statistics for Spatial Data, Revised Edition*. New York: John Wiley and Sons.

Energy Information Agency. 2004. *The National Energy Modeling System: An Overview 2003*. DOE / EIA Report # 0581(2003).

Johannesson, G. and J.S. Stewart. 2005. *Geospatial Statistics and Issues in Energy Modeling. GIS/Regionalization Workshop for Energy Efficiency and Renewable Energy*. May 10 – 11. Arlington RAND Offices. http://www.nrel.gov/analysis/workshops/gis_workshop_05.html

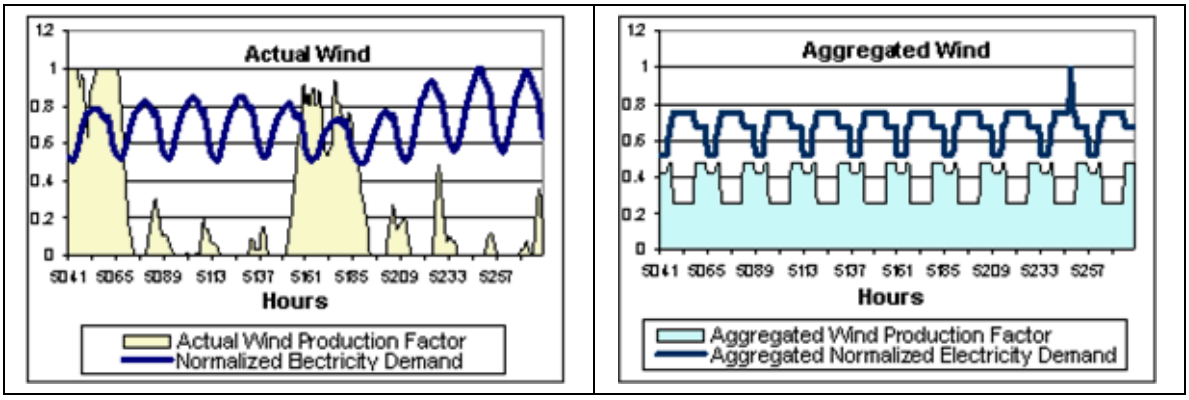
Johannesson, G., J.S. Stewart, C.D. Barr, L.B. Sabeff, R. George, D. Heimiller, A. Milbrandt. 2006. *Spatial Statistical Procedures to Validate Input Data in Energy Models*. UCRL Technical Report # 218702.

Kenny, J.F. eds. 2004. *Guidelines for Preparation of State Water-use Estimates*. USGS Techniques and Methods # 4-A4.

Lamont, A. and T. Wu. 2005. *Impact of Time Resolution on the Projected Rates of Market Penetration by Intermittent Generation Technologies*. Manuscript.

Schabenberger, O. and C.A. Gotway. 2005. *Statistical Methods for Spatial Data Analysis*. Boca Raton: Chapmand and Hall/CRC.

Images:



(a) (b)

Figure 1. (a) 10-day actual wind production and normalized electricity demand. (b) 10-day aggregated wind production and normalized electricity demand.

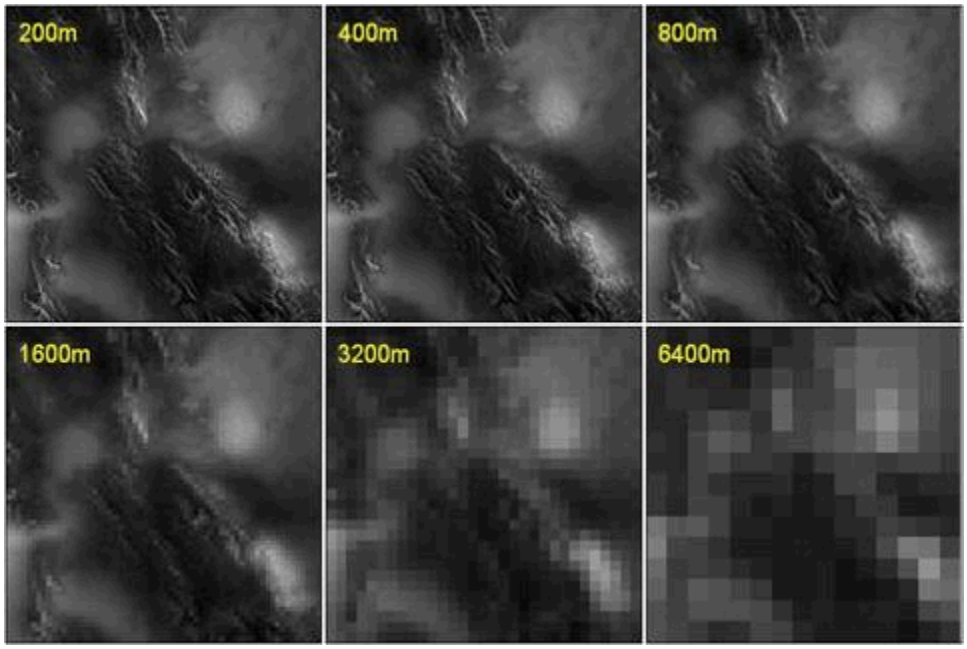


Figure 2. Average annual wind density at 50-m height at different resolutions (dark is low, white is high).

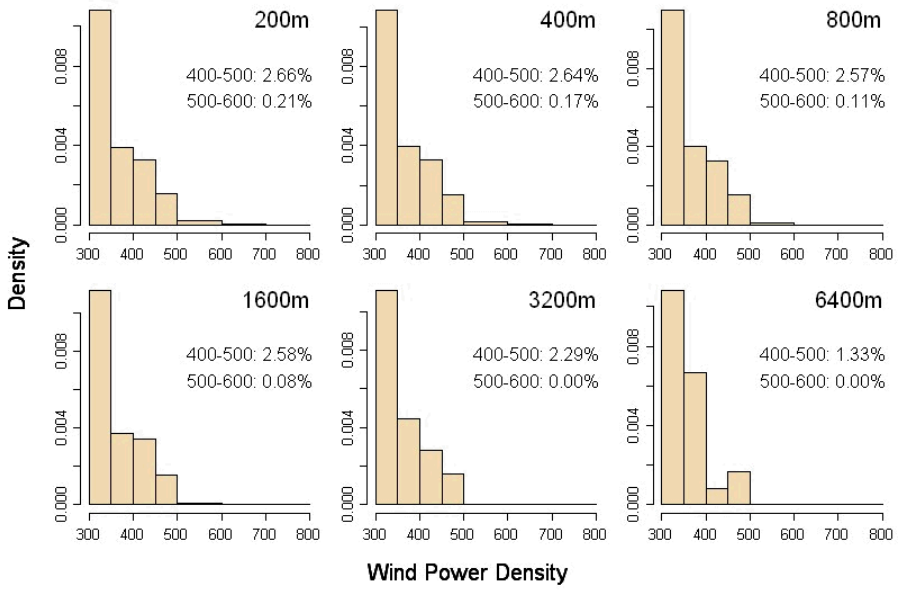


Figure 3. Histogram for the wind-power density reported at the six different resolutions..

(a) Residue Biomass (b) Residue Density (c) Land Use

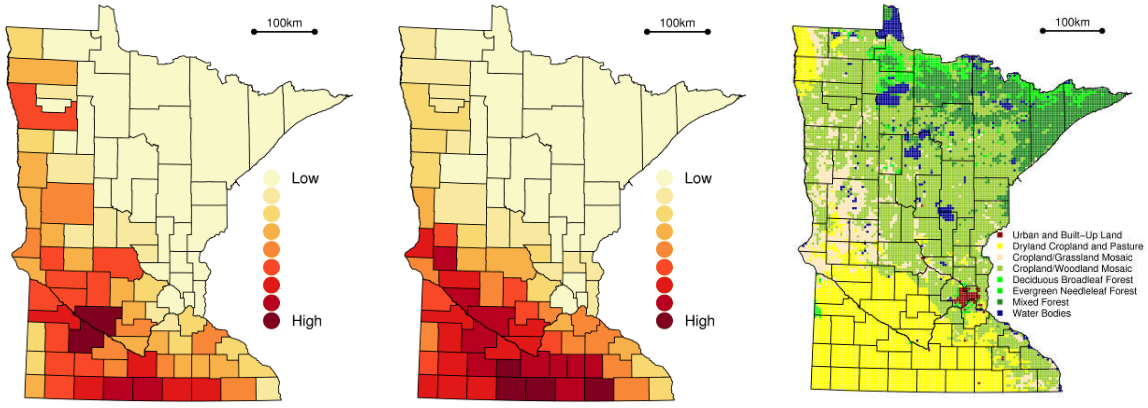


Figure 4. Minnesota annual crop residue biomass (a) and density (b) by county, along with land-use pattern (c).

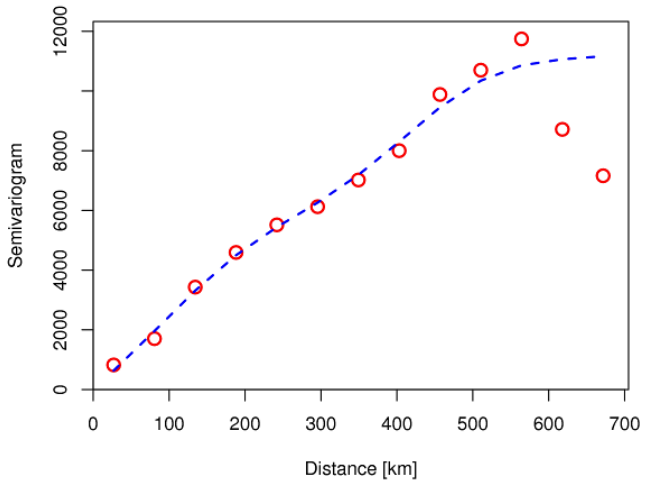
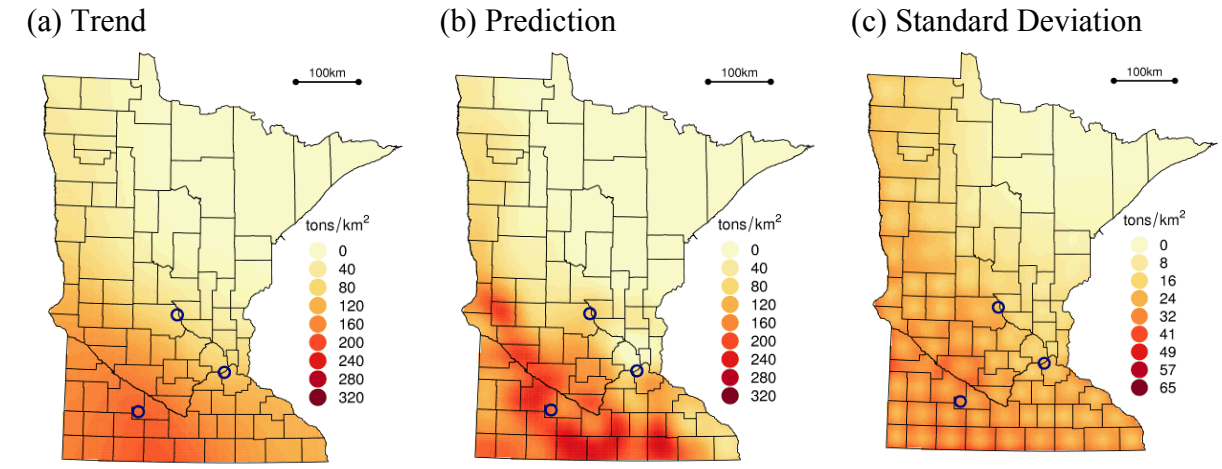


Figure 5. Semivariogram of county residue biodensity (right) and the square-root of the semivariogram. Distances are based on county centroids, and the density is in tonnes/km^2.



.

Figure 6. Residues density results from a spatial model with a smooth trend:(a) large-scale residue density trend, (b) residue density prediction, (c) the marginal prediction uncertainty

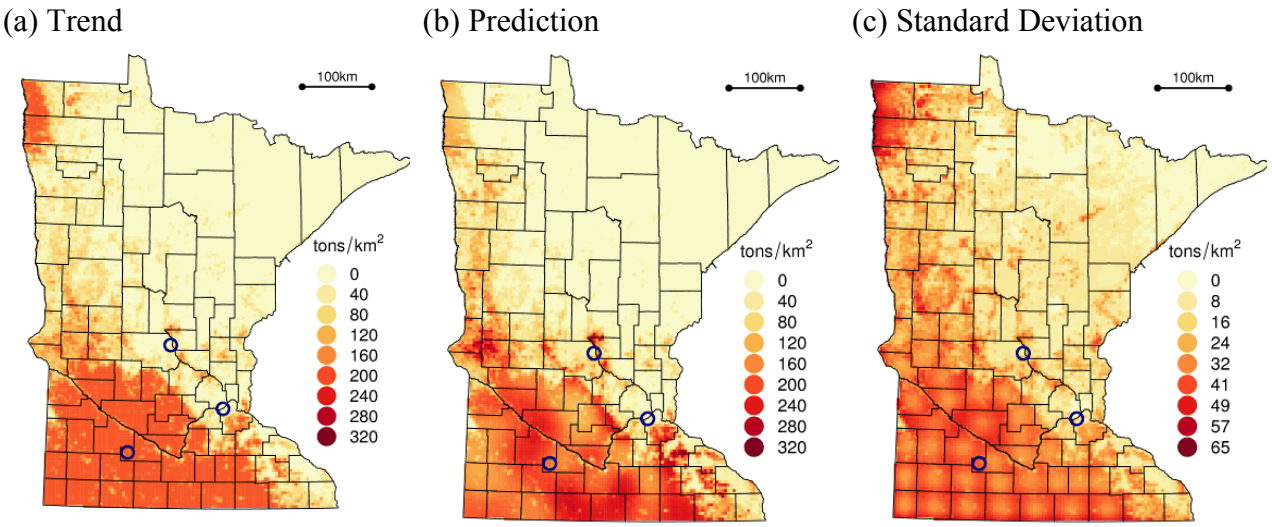


Figure 7. Residue density results from a spatial model taking advantage of land-use data:(a) land-use-based large-scale residue density trend, (b) residue density prediction, and (c) the marginal prediction uncertainty.

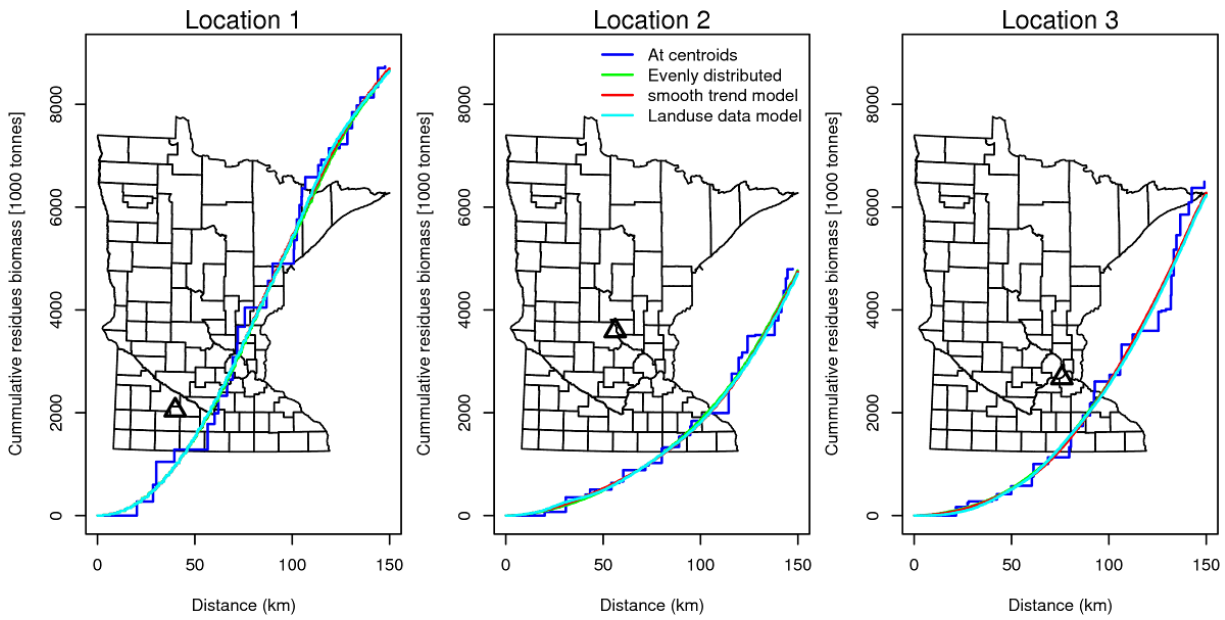


Figure 8. The cumulative residue biomass within a given (air) distance from three selected locations from four different residue biomass models (legend).